Statistical Data Mining (half question)

6. Suppose we train the following three binary classifiers.

1. LDA: The class conditional densities are Gaussian and admit the same covariance matrix, i.e. $p(\mathbf{x}|\, Y = c) = \mathcal{N}(\mathbf{x}; \mu_c, \mathbf{\Sigma})$. We assume $P(Y = c) = \frac{1}{2}$.

2. LinLog: A logistic regression model with linear features.

3. QuadLog: A logistic regression model using linear and quadratic features (i.e., polynomial basis function expansion of degree 2).

We estimate the parameters of the models by maximizing the associated likelihood function for LDA or conditional likelihood function for LinLog and QuadLog. After training we compute the performance of each model M on the training set as follows:

$$l(M) = \frac{1}{n} \sum_{i=1}^{n} \log\ P\left(y^i \middle|\, \mathbf{x}^i, \widehat{\theta}, M\right)$$

where $\widehat{\theta}$ is the parameter estimate of the unknown parameters for model M; e.g. $\theta = (\mu_1, \mu_2, \mathbf{\Sigma})$ for M =LDA.

We now want to compare the performance of each model. We will write $M \leqslant M'$ if we have $l(M) \leqslant l(M')$ for *any* training set. For each of the following model pairs, state whether $M \leqslant M'$, $M \geqslant M'$, or whether no such statement can be made (i.e., M might sometimes be better than M' and sometimes worse); also, for each question, briefly (1-2 sentences) explain why.

(a) LDA, LinLog.

(b) LinLog, QuadLog.

Now suppose we measure performance in terms of the average misclassification rate on the training set:

$$R(M) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(y^i \neq \widehat{y}\left(\mathbf{x}^i\right)\right).$$

(c) Is it true in general that $M > M'$ implies that $R(M) < R(M')$? Explain why or why not.

Statistical Data Mining (half question)

7. In binary classification, the loss function we usually want to minimize is the risk associated with the 0/1 loss:

$$L\left(y, \widehat{y}\left(x\right)\right) = \mathbb{I}\left(\widehat{y}\left(x\right) \neq y\right)$$

where $\widehat{y}\left(x\right), y \in \{0, 1\}$. In this problem we will consider the effect of using an asymmetric loss function

$$L_{\alpha,\beta}\left(y, \widehat{y}\left(x\right)\right) = \alpha\mathbb{I}\left(y = 0, \widehat{y}\left(x\right) = 1\right) + \beta\mathbb{I}\left(y = 1, \widehat{y}\left(x\right) = 0\right)$$

where $\alpha, \beta > 0$.

(a) Determine the Bayes optimal classifier, i.e. the classifier that achieves the minimum risk for the loss function $L_{\alpha,\beta}$.

(b) Suppose $P\left(Y = 0\right)$ is very small. This means that the classifier $\widehat{y}\left(x\right) = 1$ for all x will have a small risk under the 0/1 loss function. We may try to put the two classes on even footing by considering the risk

$$R = P\left(\widehat{y}\left(X\right) = 1 | \, Y = 0\right) + P\left(\widehat{y}\left(X\right) = 0 | \, Y = 1\right)$$

Show that this risk can be rewritten as the expected loss function $L_{\alpha,\beta}$ for certain values of α, β.